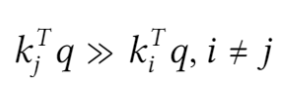
1.a)i)

α is kind of a categorical probability distribution because:

n number of categories, α1, . . . , αn event probabilities

and αi > 0, which are the prerequisites for a categorical probability distribution.

ii)

copying in attention

iii)

we know that αi > 0,ans also from αj = we understand that

αj = 1 so = vj aj = vj .

iv)

When one of the key vectors is almost identical to the given query, the attention weight assigned to it will be significantly higher. As a result, the output will be strongly influenced by that specific key's value. In this way, we can say that the model has copied the value.

b)i)

M(Va + Vb) = Va

M Va = M(c1a1 + c2a2 + …. + cmam) = MAc =Va  and aj T ai = 0 if i==j aj Tai = 1

* M = AT
* A⊤Ac = c1a1 ⊤a1 + · · · + cmam⊤am = c

M Vb = M(d1b1 + d2b2 + …. + dpbp) = MBd =Vb  and aj Tbk = 0

* M = AT
* A⊤Bd = d1a1⊤b1 + · · · + dpap⊤bm = 0

Mva + Mvb = va =>  A⊤Ac + A⊤Bd = c + 0 = c

So M = AT  and vector of weights is c.

ii)

c = (1/2)( Va + Vb) => xa= xb = 0.5 and part a

* q⊤ka = q⊤kb ≫q⊤ki ∀i != a, b
* q⊤ka = q⊤kb = β => = for β ≫ 0
* exp(β) = ∞ =>
* q = β(ka + kb) β ≫ 0

c)i)

Considering that α tends to approach zero, the covariance matrices' diagonal elements also become extremely small. so, when sampling ki with mean µi and covariance Σi, the sampled value ki will have a value close to µi.

Furthermore, since the means µi are all perpendicular it leads to the same expression for q, which is β(µa + µb).

ii)

we know µi T µi = 1 and α is small and αI + (1/2) (µaTµa )

* ka ∈ [0.5µa, 1.5µa] i != a
* ka = Xµa and X = N (1, 0.5)
* ki = µi ∀I != a
* ka T q = X µa T β(µa + µb) = Xβ β ≫ 0
* Kb T q ≈ µb T β(µa + µb) = β β ≫ 0
* Ki T q ≈ µi T β(µa + µb) = β(µi T µa + µi T µb) = β(0 + 0) = 0 β ≫ 0
* = =
* = =
* X is a shifted softmax to right by 1. So for the start and end of X period which is [0.5 , 1.5]
* For X = 0.5 and β ≫ 0 = = 0 and = = 1
* = = 1 and = = 0

So c = va if X is 1.5 and c = vb if X is 0.5

It differs from part i because in part i c is balanced combination of both va and vb but here c swings between these.

d)i)

c1 = 1/2 (va + vb) and q1 = β(µa + µb)

c2 = 1/2 (va + vb) and q2 = β(µa + µb)

c = 1/2 (c1 + c2) = 1/2 ( 1/2 (va + vb) + 1/2 (va + vb)) = 1/4 (va + vb) + 1/4 (va + vb) = 1/2 (va + vb)

ii)

c is like part (c)ii). If we add more attention the swings of c between va and vb will be less. If X is 1 then we have :

c = va + vb = (va + vb)

2.

d) accuracy on the dev set: 10.0 out of 500.0: 2.0%

accuracy “London” : 25.0 out of 500.0: 5.0%

f) accuracy on the dev set: 72.0 out of 500.0: 14.399999999999999%

g)i) accuracy : 43.0 out of 500.0: 8.6%

ii) the complexity of attention operation is reduced to O(d × m).

complexity of self attentions in the latent transformer blocks will reduce to O(m2 ).

multi-headed attention has a time complexity of O(ℓ 2d + ℓd2 )

the time complexity of the perceiver model is O(dm + Lm2 ),

3.

a) it had basic knowledge. Because it was trained on a large dataset.

b) the information made, will be wrong and this can lead to insecure events.

c) same as part b